nature portfolio

Corresponding author(s): Lee A.D. Cooper

Last updated by author(s): Oct 12, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	Code is publicly available at: github.com/PathologyDataScience/HiPS. Processing of histology images was performed using HistomicsTK (v.1.2.10, github.com/DigitalSlideArchive/HistomicsTK), histolab (v.0.6.0, github.com/histolab/histolab), and scikit-image (v.0.18.0, scikit-image.org). Analysis of clinical outcomes data was performed using Lifelines (v.0.27.8, github.com/CamDavidsonPilon/lifelines). Enrichment analysis with RNA profiles was performed using GSEAPy (v.1.0.6, gseapy.rtfd.io). Additional python libraries used for database management, graphical plotting, scientific calculations, and other tasks includes: numpy v.1.19.4, pandas v.1.1.5, SQLAlchemy v.1.3.21, scipy v.1.5.4, scikit-learn v.0.23.2, imageio v.2.9.0, pillow v.8.0.1, matplotlib v.3.3.3, seaborn v.0.11.0, torch v.1.7.1, torchvision v.0.8.2, and pyvips v.2.1.15.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Table S27 contains our calculated histomic feature values, HiPS scores and subscores, and related data for the TCGA cohort. We provide this to facilitate reproducibility and to act as a resource for the scientific community. TCGA clinical data and WSIs are publicly available at: gdc.cancer.gov. The Breast Cancer Semantic Segmentation dataset is available at: github.com/PathologyDataScience/BCSS, while the NuCLS dataset is available at: sites.google.com/view/nucls. These datasets were combined to produce the PanopTILs dataset, available at: sites.google.com/view/panoptils. Requests for American Cancer Society data from the CPS-II or CPS-3 studies should be submitted to maddison.hall@cancer.org. Requests PLCO data should be submitted at: cdas.cancer.gov/learn/plco. Breast cancer genomic subtypes, hypoxia scores, fraction genome altered, aneuploidy scores, and mRNA expression profiles were obtained from the Genomic Data Commons Pancancer Atlas: gdc.cancer.gov/about-data/publications/pancanatlas. Immune subtypes and related pathway activations, as well as angiogenesis and lymphangiogenesis scores were obtained from the PanImmune dataset: gdc.cancer.gov/about-data/publications/panimmune. CAF subtype abundance data for TCGA was obtained from: Li, B. et al. Cell-type deconvolution analysis identifies cancer-associated myofibroblast component as a poor prognostic factor in multiple cancer types. Oncogene 40, 4686–4694 (2021). xCell cell type abundance data for TCGA was obtained from: Aran, D., Hu, Z. & Butte, A. J. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol 18, 220 (2017).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Not applicable.
Reporting on race, ethnicity, or other socially relevant groupings	Provided in Table S3.
Population characteristics	Provided in Tables S1-4.
Recruitment	- TCGA dataset: Participant recruitment details can be accessed through the National Cancer Institute Genomic Data Commons (GDC) portal: https://gdc.cancer.gov/. - The CPS studies are prospective observational cohort studies of cancer risk factors, morbidity, and mortality that the American Cancer Society conducted. Deidentified clinical and imaging data was shared with Emory University and Northwestern University through data sharing agreements. Adult men and women with no personal history of cancer were enrolled in 1982 in CPS-II and 2006-2013 in CPS-3. CPS-II participants enrolled in a subcohort followed for cancer incidence and mortality starting in 1992/1993. CPS-II cancer incidence follow-up is complete through 2017 (mortality through 12/31/2018). In CPS-3, follow-up is ongoing but currently complete through 2018 (mortality through 2017). We included female participants who developed breast cancer and for whom tissue slides were available for scanning. The included participants were diagnosed between the start of follow-up (CPS-II: 1992-1993, CPS-3: 2006-2013) and the last administrative date (CPS-II: 2017, CPS-3: 2018). Details are provided in the following publications: |_ Calle, E. E. et al. The American Cancer Society Cancer Prevention Study II Nutrition Cohort: rationale, study design, and baseline characteristics. Cancer 94, 2490–501 (2002). |_ Patel, A. V et al. The American Cancer Society's Cancer Prevention Study 3 (CPS-3): Recruitment, study design, and baseline characteristics. |_ Garfinkel, L. Selection, follow-up, and analysis in the American Cancer Society prospective studies. Natl Cancer Inst Monogr 67, 49–52 (1985). |_ Stellman, S. D. & Garfinkel, L. Smoking habits and tar levels in a new American Cancer Society prospective study of 1.2 million men and women. J Natl Cancer Inst 76, 1057–63 (1986). Cancer 123, 2014–2024 (2017). - The Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial was designed to evaluate the efficacy of specific screening procedures in reducing mortality rates associated with prostate, lung, colorectal, and ovarian cancers. This PLCO trial adopted a randomized, controlled design and recruited participants from 1993 to 2001. Cancer incidence data were amassed until December 31, 2009, while mortality information was collected through 2015. PLCO female participants who developed breast cancer after enrollment in the trial were included in our study. Deidentified clinical and WSI data were obtained with permission from the National Cancer Institute Cancer Data Access System. Details are provided in the following publication: Zhu, C. S. et al. The Prostate, Lung, Colorectal and Ovarian Cancer (PLCO) Screening Trial Pathology Tissue Resource. Cancer Epidemiol Biomarkers Prev 25, 1635–1642 (2016).
Ethics oversight	All study participants participated voluntarily and provided informed consent. CPS-II and CPS-3 data sharing was approved through the Emory University Institutional Review Board, approval number IRB00045780 and IRB00059007. Deidentified CPS-II and CPS-3 data were obtained through a data sharing agreement between the American Cancer Society, Emory University, and Northwestern University. Deidentified PLCO data was obtained with permission from the National Cancer Institute Cancer Data Access System.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We used deidentified data from preexisting studies that we did not design. Details regarding sample size choice are published with each of the publicly available datasets used, including The Cancer Genome Atlas (https://www.cancer.gov/ccg/research/genome-sequencing/tcga), Cancer Prevention Studies (https://www.cancer.org/research/population-science/cancer-prevention-and-survivorship-research-team/acs-cancer-prevention-studies.html) and the PLCO trial (https://cdas.cancer.gov/plco/).
Data exclusions	All non-metastatic breast cancer cases were included in the analysis.
Replication	- We provide all code needed to replicate this study at: github.com/PathologyDataScience/HiPS - To replicate TCGA cohort findings, we provide Table S27 which contains our calculated histomic feature values, HiPS scores and subscores, and related data for the TCGA cohort. - Requests for CPS-II or CPS-3 data should be submitted to the ACS. - Requests PLCO data should be submitted at: https://cdas.cancer.gov/learn/plco/.
Randomization	Not applicable because we used retrospective data from preexisting studies that we did not design. Our study is an exploratory study, not a prospective clinical trial.
Blinding	Not applicable, because our study is an exploratory study based on prior published retrospective data, not a prospective clinical trial.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	Not applicable. We used deidentified clinical data from the preexisting TCGA, CPS-II, CPS-3, and PLCO studies.
Study protocol	Not applicable. We used deidentified clinical data from the preexisting TCGA, CPS-II, CPS-3, and PLCO studies.
Data collection	Not applicable. We used deidentified clinical data from the preexisting TCGA, CPS-II, CPS-3, and PLCO studies.
Outcomes	Not applicable. We used deidentified clinical data from the preexisting TCGA, CPS-II, CPS-3, and PLCO studies.